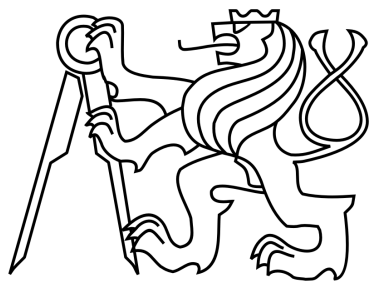




# Chybějící atributy a postupy pro jejich náhradu



# — Jedná se o součást čištění dat



Čistota dat je velmi důležitá, neboť kvalita dat zásadně ovlivňuje kvalitu výsledků, které DM vyprodukuje, neboť platí "*Garbage in, garbage out!*"

- Jaké kroky se používají při čištění dat?
  - **Řešení problému „chybějících údajů“**
  - Identifikace mimořádných hodnot (*outliers*) a vyhlazení zašuměných dat
  - Oprava nekonzistentních údajů v datech
  - Řešení redundancí vzniklých při integraci dat

Z jakých důvodů mohou data chybět?

Pacient č.	Teplota	Bolest hlavy	Nevolnost	Klasifikace (diagnoza chřipka?)
1	zvýšená	?	ne	ano
2	vysoká	ano	ano	ano
3	?	ne	ne	ne
4	zvýšená	ano	ano	ano
5	zvýšená	?	ano	ne
6	normální	ano	ne	ne
7	normální	ne	ano	ne
8	?	ano	?	ano

- Bylo zbytečné je zjišťovat, neboť rozhodnutí lze udělat i bez nich nebo v daném případě nemají smysl (např. spotřeba u bezmotorového dopravního prostředku) : data jsou pak označovány jako **údaje, o které nestojíme** („do not care“)
- Jedná se o **ztracené údaje, tedy** ty hodnoty, které
  - se nepodařilo zjistit z neznámých důvodů,
  - byly omylem vymazány,
  - šlo o evidentní překlep, ....

# Způsoby řešení problému chybějících atributů



- ❖ **Sekvenční metody** jsou součástí předzpracování dat. Jejich cílem je ze souboru s nekompletními daty vytvořit **nový soubor, který je úplný**: příslušné případy mohou být vynechány nebo naopak v nich chybějící data doplněna.
- ❖ **Paralelní metody** řeší problém chybějících hodnot až v průběhu nasazení algoritmu strojového učení (např. při hledání příslušných rozhodovacích pravidel)



# SEKVENČNÍ METODY

# 1. Vynechání neúplných příkladů



- ❖ Nejjednodušší řešení.
- ❖ Ovšem tím se připravíme o významnou část cenných dat: z původní množiny případů může zůstat malý díl (zde jen polovina!)

Pacient č.	Teplota	Bolest hlavy	Nevolnost	Klasifikace (diagnoza chřipka?)
2	vysoká	ano	ano	ano
4	zvýšená	ano	ano	ano
6	normální	ano	ne	ne
7	normální	ne	ano	ne

## 2a. Náhrada nejobvyklejší hodnotou



Pro všechny atributy se zjistí relativní frekvence jednotlivých hodnot a pro náhradu se zvolí hodnota s nejvyšší frekvencí, viz příklad:

- ❖ Teplota: **zvýšená** (3/8), normální (2/8), vysoká (1/8)
- ❖ Bolest\_hlavy: **ano** (4/8), ne (2/8)
- ❖ Nevolnost: **ano** (4/8), ne (3/8)

Pacient č.	Teplota	Bolest hlavy	Nevolnost	Klasifikace (chřipka?)
1	zvýšená	?	ne	ano
2	vysoká	ano	ano	ano
3	?	ne	ne	ne
4	zvýšená	ano	ano	ano
5	zvýšená	?	ano	ne
6	normální	ano	ne	ne
7	normální	ne	ano	ne
8	?	ano	?	ano

Pacient č.	Teplota	Bolest hlavy	Nevolnost	Klasifikace (chřipka?)
1	zvýšená	ano	ne	ano
2	vysoká	ano	ano	ano
3	zvýšená	ne	ne	ne
4	zvýšená	ano	ano	ano
5	zvýšená	anp	ano	ne
6	normální	ano	ne	ne
7	normální	ne	ano	ne
8	zvýšená	ano	ano	ano

**Výsledek náhrady**

Přístup použitý např. v algoritmu CN2

## 2b. Náhrada nejobvyklejší hodnotou **vzhledem ke klasifikaci případu**



Při nahrazování hodnoty pro případ, který je klasifikován do třídy „c1“, se vybírá nejméně frekventovaná odpověď zjištěná **jenom** mezi případy dané třídy „c1“.

Např. pro pacienty č. 3 a 5 zařazené do třídy „diagnóza chřipka ne“ musíme zjistit frekvence výskytu hodnot mezi pacienty {3,5,6,7}:

- ❖ Teplota: zvýšená (1/4), **normální** (2/4), vysoká (0/4)
- ❖ Bolest\_hlavy: ano (1/4), **ne** (2/4)

Tato náhrada není použita pro pacienty č. 1 a 8, pro něž musíme naopak zjišťovat relativní frekvenci výskytů v množině pacientů {1,2,4,8}, tedy:

- ❖ Bolest\_hlavy: **ano** (3/4), ne (0/4)
- ❖ Nevolnost: **ano** (2/4), ne (1/4)

### Výsledek náhrady

Použito např. v ASSISTANT (Kononenko et. al., 1984)

Pacient č.	Teplota	Bolest hlavy	Nevolnost	Klasifikace (chřipka?)
1	zvýšená	?	Ne	ano
2	vysoká	ano	Ano	ano
3	?	ne	Ne	ne
4	zvýšená	ano	ano	ano
5	zvýšená	?	ano	ne
6	normální	ano	ne	ne
7	normální	ne	ano	ne
8	?	ano	?	ano

Pacient č.	Teplota	Bolest hlavy	Nevolnost	Klasifikace (chřipka?)
1	zvýšená	ano	ne	ano
2	vysoká	Ano	ano	ano
3	normální	ne	ne	ne
4	zvýšená	ano	ano	ano
5	zvýšená	ne	ano	ne
6	normální	ano	ne	ne
7	normální	ne	ano	ne
8	zvýšená	ano	ano	ano



# 3. Náhrada všemi možnými hodnotami daného atributu

Každý případ (pacient) je popsán více alternativními vektory údajů. Výsledná data mohou být nekonzistentní (např. řádky 1'' a 3') - vhodné jen pro metody stroj. učení, které s tím dovedou pracovat (např. užívající rough sets theory). Opět jsou 2 přístupy:

❖ každá možná hodnota chybějícího atributu je použita



❖ každá možná hodnota chybějícího atributu **v dané klasifikační třídě** je použita:



Např. u pacienta 1 tedy není použita náhrada hodnotou „ne“, která pro třídu „diagnóza chřipka ano“, tj. pro pacienty { 1,2,4,8 } není nikdy v datech použita.

Dále u pacienta 3 tedy není použita náhrada hodnotou „vysoká“, která pro třídu „diagnóza chřipka ne“, tj. pro pacienty {3,5,6,7} není nikdy v datech použita.

Pacient č.	Teplota	Bolest hlavy	Nevolnost	Klasifikace (chřipka?)
1	zvýšená	?	ne	ano
2	vysoká	ano	ano	ano
3	?	ne	ne	ne
4	zvýšená	ano	ano	ano
5	zvýšená	?	ano	ne
6	normální	ano	ne	ne
7	normální	ne	ano	ne
8	?	ano	?	ano

Pacient č.	Teplota	Bolest hlavy	Nevolnost	Klasifikace (chřipka?)
1'	zvýšená	ano	ne	ano
1''	zvýšená	ne	ne	ano
2	vysoká	ano	ano	ano
3'	zvýšená	ne	ne	ne
3''	vysoká	ne	ne	ne
3'''	normální	ne	ne	ne
...				

Pacient č.	Teplota	Bolest hlavy	Nevolnost	Klasifikace (chřipka?)
1'	zvýšená	ano	ne	ano
2	vysoká	ano	ano	ano
3'	zvýšená	ne	ne	ne
3'''	normální	ne	ne	ne
...				

## 4. Náhrada průměrem (u spoj.dat) nebo nejčastější hodnotou

Průměrná hodnota pro teplotu je

37,4° C pro celý soubor.

37,9° C pro data klasifikovaná „diagnóza chřipka ano“,

36,4° C pro „diagnóza chřipka ne“.

Opět jsou 2 přístupy:

❖ Náhrada je provedena pomocí **průměru na celém souboru**

❖ Pro náhradu se používají hodnoty **průměru v příslušné klasifikační třídě.**

Pacient č.	Teplota	Bolest hlavy	Nevolnost	Klasifikace (chřipka?)
1	37,5	?	ne	ano
2	39,1	ano	ano	ano
3	?	ne	ne	ne
4	37,2	ano	ano	ano
5	37,6	?	ano	ne
6	36,2	ano	ne	ne
7	36,7	ne	ano	Ne
8	?	ano	?	ano

Pacient č.	Teplota	Bolest hlavy	Nevolnost	Klasifikace (chřipka?)
1	37,5	ano	ne	ano
2	39,1	ano	ano	ano
3	37,4	ne	ne	ne
4	37,2	ano	ano	ano
5	37,6	ano	ano	ne
...				

Pacient č.	Teplota	Bolest hlavy	Nevolnost	Klasifikace (chřipka?)
1	37,5	ano	ne	ano
2	39,1	ano	ano	ano
3	36,4	ne	ne	ne
4	37,2	ano	ano	ano
5	37,6	ano	ano	ne
...				

# 5. Náhrada hodnotou nejbližšího souseda

Vzdálenost 2 vzorků  $x$  a  $y$  popsaných  $n$  atributy, jejichž některé hodnoty mohou chybět:

$$distance(x, y) = \sum_{i=1}^n distance(x_i, y_i), \text{ kde}$$

$$distance(x_i, y_i) = \begin{cases} 0 & \text{pokud } x_i = y_i \\ 1 & \text{pokud hodnoty } x_i \text{ a } y_i \text{ jsou výčtového typu nebo jedna z nich chybí} \\ |x_i - y_i| / r_i & \text{jinak (tedy jde o reálné hodnoty s rozsahem } r_i, \text{ odpovídajícím rozdílu mezi max a min hodnotou)} \end{cases}$$

Pacient č.	Teplota	Bolest hlavy	Nevolnost	Klasifikace (chřipka?)
1	37,5	?	ne	ano
2	39,1	ano	ano	ano
3	?	ne	ne	ne
4	37,2	ano	ano	ano
5	37,6	?	ano	ne
6	36,1	ano	ne	ne
7	36,7	ne	ano	ne
8	?	ano	?	ano

a) Postup náhrady hodnoty atributu „**bolest hlavy**“ pro vzorek **1** přes všechna data:

d(1,2)	d(1,3)	d(1,4)	d(1,5)	d(1,6)	d(1,7)	d(1,8)

1. Vypočti vzdálenosti vzorku 1 od ostatních.
2. Vyber nejbližší (je jím ???)
3. Jeho hodnotu (???) použij jako náhradu pro vzorek 1.

Pacient č.	Teplota	Bolest hlavy	Nevolnost	Klasifikace (chřipka?)
1	37,5	ano	ne	ano
2	39,1	ano	ano	ano
3	36,4	ne	ne	ne
4	37,2	ano	ano	ano
...				

b) **Obdobně lze postupovat také jen uvnitř jediné třídy!**



# PARALELNÍ METODY